DOCUMENT RESUME

ED 173 396 TM 009 399

TITLE College Outcome Measures Project. COMP Annual Report, 1978.
INSTITUTION American Coll. Testing Program, Iowa City, Iowa.
SPONS AGENCY Fund for the Improvement of Postsecondary Education (DHEW), Washington, D.C.
PUB DATE 78
NOTE 36p.
AVAILABLE FROM College Outcome Measures Project, The American College Testing Program, P.O. Box 168, Iowa City, Iowa 52240 (no charge)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adults; *Basic Skills; College Students; Cost Effectiveness; Educational Assessment; *Equivalency Tests; *General Education; Minimum Competency Testing; Postsecondary Education; Program Development; *Research Projects; Student Evaluation; *Test Construction; *Testing Programs; Test Reliability; Test Validity
IDENTIFIERS *College Outcome Measures Project; Test Batteries

ABSTRACT

The College Outcome Measures Project (COMP) was organized to design, develop, validate, and disseminate assessment instruments and procedures to measure and evaluate the knowledge and skills that undergraduate students are expected to acquire as a result of general or liberal education programs, as well as those important to effective functioning in adult society. Six interrelated content and process skills were identified: (1) functioning within social institutions; (2) using science and technology; (3) using the arts; (4) communicating; (5) solving problems; and (6) clarifying social values. Major research projects were conducted on the three COMP instruments (the open-ended Measurement Battery, the Objective Test, and the Activity Inventory) with college freshmen, college seniors, postsecondary vocational-technical students, and mature adults. An advisory panel report evaluated the results of these projects, concluding that the COMP instruments have achieved reasonable validity, reliability, and cost effectiveness. Furthermore, these field tests indicated that the six skills are affected by college education, and do correlate significantly with effectively functioning adults. (GDC/CP)

COMP
Annual
College Outcom

ssessment of Gener
he American College

COMP
Annual Report 1978

College Outcome Measures Project

Aubrey Forrest, Director

Joe M. Steele, Research Psychologist

"It will always be possible to debate as to what outcomes colleges should be trying to generate in their general education programs, what priorities the different objectives should have, which of those such as here addressed can be 'taught deliberately,' and like questions. In this instance a sizable group of colleges, which are known to represent interests shared by a larger company, had the opportunity to try to clarify what they most valued among the possible outcomes to be sought and assessed. **Given the current challenge as to whether college education is really worthwhile and the extensive pressures to answer that question primarily in terms of job- and career-related concerns, the choice made by these colleges and applied in this project is a significant effort to clarify and revitalize the commitment to complementary values. It reaffirms the idea that college education should be judged by the quality of life to which it contributes and that this quality is a matter of the development of interests and concerns and of behavior applying those concerns in participation in the adult community."**

— Morris Keeton
Excerpt from Advisory and Evaluation Panel Report

Individuals who seek further information about COMP, who wish to request copies of COMP assessment instruments, or who wish to have their names added to the COMP mailing list for periodic information about the project, should write to COMP, The American College Testing Program, P.O. Box 168, Iowa City, Iowa 52240, or telephone Dr. Aubrey Forrest at 319/356-3933.

Additional copies of this 1978 Annual Report are available at no charge.

Supported by a grant from the Fund for the Improvement of Postsecondary Education, Department of Health, Education, and Welfare.

CONTENTS

For additional information about the College Outcome Measures Project (COMP), not contained in this document, refer to:

- 1977 Annual Report
- COMP Prospectus for 1978-79

INTRODUCTION

Purpose

The American College Testing Program (ACT) organized the College Outcome Measures Project (COMP) in 1976, as a major cooperative effort involving a variety of postsecondary institutions and agencies. Primary financial support for COMP has come from the Fund for the Improvement of Postsecondary Education (FIPSE). The purpose of COMP is to design, develop, validate, and implement assessment instruments and procedures to measure and evaluate the knowledge and skills (1) that undergraduate students are expected to acquire as a result of general or liberal education programs, and (2) that are important to effective functioning in adult society.

Assumptions

The first major step in the project was to identify, by sorting and analyzing the large number of goals colleges list for their programs, a set of competencies or outcome domains commonly held by the participating institutions and agencies. In attempting to define the major areas of general education knowledge and skills for which no adequate tests existed, COMP made a number of fundamental assumptions:

- COMP should assess generic college outcomes rather than discipline- or content-based outcomes.
- COMP should assess the college outcomes which postsecondary institutions usually expect to result from the general education components of their curriculums.
- COMP should assess the college outcomes necessary for effective functioning in adult roles.
- COMP should allow students the opportunity to demonstrate the full range of general knowledge and skills they possess.
- COMP should provide an assessment package which possesses face validity for students, faculty, and external publics.

Outcomes

After extensive definition and review by the twelve diverse institutions and agencies participating in COMP, six areas were identified: communicating, solving problems, clarifying values, functioning within social institutions, using science and technology, and using the arts. Participants in the project believe that this list contains some, but not all, of the competencies or outcomes that should result from general or liberal education and that should be important to effective adult functioning. Much of the research of the project centers around an empirical examination of these beliefs. The major general knowledge and skill areas upon which the project is focused are:

Communicating: Can send and receive information in a variety of modes (written, graphic, oral, numeric, and symbolic), within a variety of settings (one-to-one, in small and large groups), and for a variety of purposes (for example, to inform, to understand, to persuade, and to analyze).

Solving Problems: Can analyze a variety of problems (for example, scientific, social, personal); select or create solutions to problems; and implement solutions.

Clarifying Values: Can identify one's personal values and the personal values of other individuals; understand how personal values develop; and analyze the implications of decisions made on the basis of personally held values.

Functioning within Social Institutions: Can identify those activities and institutions which constitute the social aspects of a culture (for example, governmental and economic systems, religion, marital and familial institutions, employment, and civic, volunteer, and recreational organizations); understand the impact that social institutions have on individuals in a culture; and analyze one's own and others' personal functioning within social institutions.

Using Science and Technology: Can identify those activities and products which constitute the scientific/technological aspects of a culture (for example, transportation, housing, energy, processed food, clothing, health maintenance, entertainment and recreation, mood-altering, national defense, communication, and data processing); understand the impact of such activities and products on the individuals and the physical environment in a culture; and analyze the uses of technological products in a culture and one's personal use of such products.

Using the Arts: Can identify those activities and products which constitute the artistic aspects of a culture (for example, graphic art, music, drama, literature, dance, sculpture, film, architecture); understand the impact that art, in its various forms, has on individuals in a culture; analyze uses of works of art within a culture and one's personal use of art.

Assessment Design

In its first two years of research and development, COMP has initiated three different approaches to the assessment of the six general education outcome areas that were defined. The key features of this assessment package were determined in accordance with assumptions about the assessment package and the nature of the knowledge and skills to be assessed. These key features are:

- the employment of realistic stimulus material drawn from the adult public domain, such as television documentaries, articles, advertisements, short stories, charts and diagrams from magazines and newspapers; letters to editors; business memoranda; discussions; telephone calls; speeches; art posters and prints; musical recordings; films;
- the use of an open-response format (in addition to the usual multiple-choice questions) which requires students to provide their own answers rather than simply selecting one of several answers supplied to them;
- development of questions that require the application of knowledge and skills to problems and issues commonly confronted by adults;

- techniques for the evaluation of student responses which allow faculty to base their judgments on standardized rating scales;
- a modular assessment package from which institutions may select those portions of the materials which meet their needs;
- multiple perspectives from which to assess each student response in order to yield judgments about at least two areas of knowledge and skill. For example, a student's writing sample might be assessed to determine both skill in solving problems and knowledge about the use of science and technology.

Considerable effort has been made to develop instruments that are content fair; process fair, and efficient. To be fair, an assessment of content should cover a broad scope of processes, just as a fair assessment of process should cover a broad scope of content. In COMP, efficiency is gained by interrelating the content- and process-related areas defined above. This interrelationship is illustrated in the following chart. The three content-related outcome domains assessed by COMP (Functioning within Social Institutions, Using Science and Technology, and Using the Arts) are shown at the top of the chart. The three process-related outcome domains (Communicating, Solving Problems, and Clarifying Values) are shown at the left side of the chart. Each of the nine areas in the chart represents the interrelationship of content and process skills. This approach to definition and assessment of general education outcomes is further made cost-effective by designing modularized instruments so that specific areas and components can be selected to fit the unique needs of a specific institution.

	Functioning within Social Institutions	**Using Science and Technology**	**Using the Arts**
Communicating	Communicating about Social Institutions	Communicating about Science and Technology	Communicating about the Arts
Solving Problems	Solving Social Problems	Solving Scientific/ Technological Problems	Solving Artistic Problems
Clarifying Values	Clarifying Social Values	Clarifying Scientific/ Technological Values	Clarifying Artistic Values

A detailed description of the assessment design, including specification of stimulus and response modes, is included in the COMP 1977 Annual Report.

COMP/ACT Assessment Instruments

The three instruments developed to date are: (1) a Measurement Battery, (2) an Objective Test, and (3) an Activity Inventory. Questions in the Measurement Battery and the Objective Test are based on television documentaries, recent magazine articles, ads, short stories, art prints, music, discussions, and newscasts. In the case of the Measurement Battery, participants view these materials and then respond with short written answers, with some longer essays written in the form of letters, and with some oral responses which are audiotaped or videotaped. In the Objective Test, participants respond to the same materials through innovative multiple-choice questions. The Activity Inventory uses a multiple choice format to assess the quality and quantity of participation in out-of-class activities related to the six outcome areas.

Detailed descriptions of these three instruments appear in the COMP Prospectus for 1978-79.

Institutional Use of COMP Materials

Increasing numbers of postsecondary institutions have come to recognize the importance of carefully specifying the general knowledge, skills, and attitudes needed by their graduates to function effectively as adults, and then accurately assessing student achievement of such general knowledge, skills, and attitudes. Institutions find they need such assessment information in order to:

- assure that students are reaching general or liberal education goals and are receiving appropriate credit and recognition for their achievement of these goals;
- shape the curriculum so as to include those learning activities which will help students obtain the knowledge, skills, and attitudes crucial to adult functioning;
- help students plan the general or liberal education programs that best meet their expectations and priorities, as well as those of society;
- assure the consumers and supporters of postsecondary education that general and liberal education programs effectively impart at least some of the general knowledge and skills students need to function effectively in adult society.

Assumptions about Use

The nontraditional nature and design of COMP materials means that specialized training is required for their use and interpretation. COMP provides institutions with a set of services to aid in defining and measuring the overall outcome of a general program of education. Specifically, these services include:

- assisting in outcome definition and examining the fit between each institution's general education outcomes and COMP assessment instruments;

- advising on appropriate and inappropriate uses and interpretations of COMP assessment materials;
- counseling in the setting of reasonable assessment standards;
- assisting in the development of supplemental assessment materials to better assess each institution's unique general education program.

An institution exploring use of COMP would be expected to:

- consider the validity of COMP with regard to its own program objectives and instructional activities;
- try out the basic COMP materials to establish local norms and facilitate rational standard setting;
- weight scores or augment areas assessed to tailor COMP to the unique needs and purposes of the institution;
- study the effectiveness of programs intended to develop the skills to be measured before using COMP for certification, crediting, or placement of individual students in these areas.

Steps in Development and Validation

The plan for development of the COMP Assessment Package involves a major commitment of time and effort by ACT staff over an 8-10-year period. A key feature of the COMP research and development effort has been the substantial assistance provided by faculty and administrators at the participating institutions and agencies in (1) the identification of outcomes, (2) the development of COMP materials, and (3) a series of field research tryouts of the assessment instruments with students and nonenrolled adults.

The initial field trials in spring 1977 tested the feasibility of the assessment approach described in this brief overview. It yielded useful information about the face validity of materials and evidence that the outcomes identified were indeed addressing skills developed in college programs. The materials yielded results not reflected in traditional measures, such as GPA, and not accounted for by maturation or ability alone.

Validity

A pilot validity study was conducted in summer 1977 with mature adults identified as functioning effectively in three areas relevant to portions of the assessment materials. This exploratory study provided support for the belief that the domains assessed by COMP are important to effective functioning in adult roles. Four more adult studies conducted in fall 1977 and spring 1978 yielded similar results. A series of comprehensive adult studies is planned to document more fully the relationship between general education outcomes and adult functioning in a variety of roles. The first of these expanded validity studies was begun in August 1978.

Reliability

Field trials of revised and newly developed assessment materials were conducted in fall 1977 and spring 1978. These studies aimed to determine the objectivity and reliability of various components of the materials, especially the degree to which faculty could agree on the evaluation of student responses. ACT staff were reassured and encouraged by the findings from these trials. In summer 1978 further studies were conducted to explore the degree to which training of faculty in the use of scales to rate student responses would yield even higher levels of agreement. The purpose of the remainder of this report is to describe these validity and reliability studies conducted in 1977-78.

The COMP Prospectus for 1978-79 describes further plans for refinement and validation of the COMP assessment package and invites participation by postsecondary institutions in experimental use of the materials. The final section of this report also contains a formal review of the first two years of COMP efforts by Morris T. Keeton, chairman of the COMP Advisory and Evaluation Panel.

Fall 1977 Field Trials of the COMP/ACT Measurement Battery, Form II

The COMP/ACT Measurement Battery, Form II, is an instrument designed to measure the ability of students to **apply**, in a variety of situations, general education skills and knowledge thought important for effective functioning in adult roles. It includes 60 questions clustered around 15 activities that use realistic stimulus materials drawn from the adult public domain (television documentaries, ads, magazine articles, music, etc.). The questions require students to provide their own answers, including nine minutes of taped responses, rather than merely selecting from among given answers. The degree of adequacy of each response is then evaluated by four trained faculty evaluators, expert in the various domains assessed, using standardized rating scales. These 60 scales allow faculty to discriminate five levels of proficiency in each student response. Administration and scoring options are available so that a college can fit the Battery to its own unique general education objectives. If the entire Battery is used, the maximum total score possible is 240 points.

In addition to a total score, seventeen subscores are derived for the COMP/ACT Measurement Battery. Table 1 shows how these subscores relate to the basic design of the instrument. The first six subscores, each based on 18 to 20 questions in the Battery, relate to the six outcome domains of general education knowledge and skills, outlined and defined on pp. 1-2. Besides the domain score for Communicating, two specific Communicating subscores, Speaking and Writing, are produced. These are defined in Table 1 and are based on nine minutes of taped responses and on sixty minutes of written responses to produce three letters. The remaining nine subscores represent the application of the skills of Communicating, Solving Problems, and Clarifying Values in social, scientific, and artistic contexts. These nine scores, each based on 6 to 8 questions in the Battery, are defined in Table 1. Thus the use of the entire Battery yields a comprehensive assessment of up to two years of an undergraduate liberal arts program. These 18 scores reflect, in sum, the ability to apply general education skills and knowledge in a wide variety of contexts.

Administration time for the entire Battery requires about six hours from students. All but nine minutes of the Battery can be administered to groups of students. The remaining nine minutes, which involve videotaping of oral responses, must be administered to one student at a time. If audiotaping is used instead of videotaping, language lab facilities can be used to tape 20-30 students at a time. It then takes about one hour per student for four faculty members to evaluate the student's written and oral responses. For example, if 100 students were tested, it would take a total of 100 person-hours for the four faculty members to complete evaluation of the 100 students. If only parts of the Battery were used, less time would be required of students and faculty.

TABLE 1

Subscores Derived from the COMP/ACT Measurement Battery

	1. FUNCTIONING WITHIN SOCIAL INSTITUTIONS	2. USING SCIENCE AND TECHNOLOGY	3. USING THE ARTS
4. COMMUNICATING	**7. Speaking:** ability to establish a human relationship with an audience, and concern about delivery (animation, fluency, poise) as well as discourse (organization of ideas, examples, detail). This score is based on nine minutes of taped responses to questions dealing with social institutions, science and technology, and art, for which advanced preparation is allowed. **8. Writing:** ability to send information to a particular audience with a focus on persuasion in which points are developed in a clear and controlled way and correct and lively use of language is demonstrated. This score is based on sixty minutes of writing to produce three letters.		
	9. Communicating about Social Institutions: ability to send and receive information (including numeric and graphic material) related to activities and institutions which constitute the social aspects of a culture.	**12. Communicating about Science and Technology:** ability to send and receive information (including numeric and graphic material) related to the scientific/technological aspects of a culture.	**15. Communicating about the Arts:** ability to send and receive information (including numeric and graphic material) related to the artistic aspects of a culture.
5. SOLVING PROBLEMS	**10. Solving Social Problems:** ability to define problems of functioning within social institutions and select approaches to solve problems, generate solutions, collect information, check logical consistency, select a good solution, and evaluate the process by which a problem was solved	**13. Solving Scientific and Technological Problems:** ability to define problems related to scientific products and the use of technology in a culture, select approaches to solve problems, generate solutions, collect information, check logical consistency, select a good solution, and evaluate the process by which a problem was solved.	**16. Solving Artistic Problems:** ability to define problems related to art and its expression and use in a culture, select approaches to solve problems, generate solutions, collect information, check logical consistency, select a good solution, and evaluate the process by which a problem was solved.
6. CLARIFYING VALUES	**11. Clarifying Social Values:** ability to identify one's own social values and those major values and issues faced by others in daily adult life in one's own and other cultures, understand how values develop, and analyze the implications of decisions made on the basis of those values	**14. Clarifying Scientific and Technological Values:** ability to identify one's own scientific/ technological values and those major values and issues faced by others in daily adult life in one's own and other cultures, understand how values develop, and analyze the implications of decisions made on the basis of those values.	**17. Clarifying Artistic Values:** ability to identify one's own artistic values and those major values and issues faced by others in daily adult life in one's own and other cultures, understand how values develop, and analyze the implications of decisions made on the basis of those values.

Seven institutions participated in the fall 1977 field test of the COMP/ACT Measurement Battery: Brigham Young University, Colgate University, Florida A & M University, Mars Hill College, Michigan State University, Our Lady of the Lake University, and the University of Nebraska-Lincoln. Each of these institutions was asked to identify 45 freshmen and 45 seniors with equal representation from the social sciences, natural sciences, and arts/humanities. All student participants were volunteers. Not all the colleges were able to fill their student quotas, with two colleges providing about one-fourth and another two about one-half the numbers requested. Faulty recording equipment and inclusion of sophomores and juniors further reduced the usable sample at two institutions. The intended and actual distribution of students is shown in Table 2. The combined sample is representative of the diversity in higher education and thus serves the purposes of the field test. Students involved at an individual institution, however, are not necessarily representative of all students at that institution.

TABLE 2

Sampling Plan and Numbers Obtained in Each Category

	Totals		Social Sciences		Natural Sciences		Arts/Humanities	
	Intended	Actual	Intended	Actual	Intended	Actual	Intended	Actual
Freshmen	315	140 (40%)	105	56 (40%)	105	45 (32%)	105	39 (28%)
Seniors	315	154 (52%)	105	64 (42%)	105	42 (27%)	105	48 (31%)
Totals		294		120 (40%)		87 (30%)		87 (30%)

Background information of the following type was requested for all students: sex, age, ACT Composite score or equivalent, and total GPA (for seniors only). A summary of these data is shown in Table 3.

TABLE 3

Student Sample Background Information

	Sex		Age				ACT Composite*				
	Men	Women	19 or below	20-22	23-30	31+	1-15	16-20	21-25	26-36	Mean
Freshmen	42%	58%	94%	5%	0	1%	15%	18%	37%	31%	22.5
Seniors	49%	57%	1%	59%	23%	16%	10%	16%	31%	43%	23.5
Totals	46%	54%	45%	33%	12%	9%	13%	17%	35%	36%	23.0

GPA (Seniors only)

0-1.00	1.01-2.00	2.01-3.00	3.01-4.00
0	4%	35%	61%

*ACT scores were available for 131 freshmen and 96 seniors

Within the total group of 294 freshmen and seniors, various subgroups were well represented. The sample was composed of 48% freshmen and 52% seniors with no significant differences in ability by grade level or major. The mean ability of the sample, based on ACT Composite or SAT Total scores, was high, above two-thirds of entering college freshmen nationally. Slightly more students represented the social sciences (40%) than natural sciences (30%) and arts/humanities (30%). Freshmen were predominantly 19 years of age or younger (94%). Ninety-nine percent of the seniors were age 20-22 (59%) or older. It was determined that all but one of the institutional groups of participants contained comparable subgroups equivalent in ability across grade level, major, sex, and age. The combined group of freshmen and seniors is judged as comparable and well-balanced. However, since each institution's sample was quite small and was probably not representative, caution is urged in interpreting results for individual institutions.

The results from the spring 1977 field test of Form I of the Battery, reported in the COMP 1977 Annual Report, were replicated and expanded by findings in the fall 1977 field trials of Form II. These results clearly reveal the discriminating power of the Battery and indicate that this type of assessment holds a great deal of promise. The following summarizes findings to date on the COMP/ACT Measurement Battery:

- The outcomes identified for the project appear to be related to the educational programs of the participating institutions and unrelated to simple maturation. The results of the first field test show a clear progression of growth across four years of college, as reported in the COMP 1977 Annual Report, p. 9. The results of the fall field test for a new form of the Battery reveal highly significant differences in performance between freshmen and senior means. Seniors scored higher than freshmen on the total Battery score and on all 17 of the subscores derived for the instrument. In both field tests there is nearly zero correlation between age and total COMP test score. No significant differences in total scores were noted by sex or major field. A study described later in this report compares college seniors in programs that include general education with vocational-technical students of comparable age and ability in programs that do not include general education. While the latter obtained substantially lower scores on the Measurement Battery than college seniors, a sizable minority could be judged as already possessing general education skills and knowledge at a collegiate level.

- The Battery appears to be measuring abilities not measured by college grades. Seniors' total scores on the Battery have shown a significant, but small relationship (.20 intercorrelation with Form I and .35 with Form II) to their college GPAs.

- The total Battery score differs from measures of general aptitude and ability. There is a clear relationship between performance on COMP and ACT Composite (or SAT Total) scores for freshmen (.67 intercorrelation with total score on the Battery). However, this level of relationship

accounts for the variance in performance of only 45% of the students. Most of the differences in performance cannot be attributed to differences in general ability.

- The Battery does seem to be measuring abilities relevant to functioning in adult roles. Adults who have been selected on the basis of significant participation in a particular community role (e.g., solving technological problems) show a strong tendency to score highest in the corresponding area of the Battery than in other areas (statistically significant beyond the .001 level). Further discussion of this finding appears in the section of this report titled "Mature Adult Performance on COMP/ACT Instruments."

- Data collected indicate that the Battery may be culture fair. The fairness of a test depends on many factors, some of which are yet to be examined. However, faculty from a variety of racial/ethnic backgrounds helped to choose stimulus materials, items, and rating scales. Local faculty, evaluating students at institutions where minorities were represented, felt materials and scales were appropriate. The students themselves, at institutions with high minority representation, responded favorably about the fairness of the Battery.

  Some limited data is also available about the performance of various groups. Freshmen at three institutions (for whom ability scores were available) were compared:

  A predominantly white institution (N = 24: all white)

  A predominantly black institution (N = 24: 23 black, 1 white)

  An institution with high minority representation (N = 24: 46% white, 50% Hispanic, 4% black)

  Mean ACT Composite scores for the three groups were not significantly different (22.96, 19.96, 19.75, respectively) and the range of scores was similar.

  Mean total scores on the Battery were essentially identical (87, 85, 86). While the mean ability of students from each of the three colleges was above the national average for their ethnic group, no one group appeared to be at an advantage or disadvantage on Battery total scores or sub-scores.

- Rating five levels of proficiency for each of the 60 Battery scales allowed some degree of success for most students. If level 5 is considered failing, 90% or more of the students provided a passing response to almost half (29) of the 60 questions. Over three-fourths of the questions were "passed" by more than 75% of the students. Of the nine questions for which fewer than 50% provided a passing response, only one could be answered by fewer than 28%. While some level of proficiency was demonstrated by most students on most items, the Battery had a high ceiling which allowed for measurement of mature, experienced adults. The average senior score on

the 18 scores of the Battery ranged from 42-53% of the maximum possible, except for two areas: Clarifying Social Values, 63%; and Solving Scientific Problems, 38%.

- Standardized rating scales enable faculty to make objective and reliable judgments about levels of performance. In the first field test (Form I of the Battery), where faculty rated responses from A-F without scales and using their own criteria, an agreement of 64% was obtained. With rating scales developed for the second field test (Form II of the Battery), 74% agreement was achieved on the pass-fail standards selected by each institution. Revised rating scales, developed in spring 1978 for readministration of Form II at two institutions, produced an agreement of 83%. These levels of agreement are quite adequate for an instrument to be used in program evaluation.

  Further refinements in scales and scoring procedures will assure levels of agreement to allow use of Battery subscores in making decisions about individual students for purposes of placement, advising, awarding credit for experiential learning, certifying for graduation, or screening for advanced training or employment. As institutions have shown a preference for making judgments of the level of proficiency rather than passing at a selected standard of performance, other statistics for determining reliability (such as the Pearson correlation) are relevant. Based on the revised rating scales for Form II, a total test score Pearson correlation of .90 was obtained. The following test-retest study reports on this information and also provides evidence of the stability of student responses.

- Faculty and student reactions to the Battery have been quite positive.

- The Battery is considered cost-effective to administer and score. At present, typical costs per student average about $25, including ACT staff time and travel for faculty training, faculty released time for scoring, administration staff time, materials preparation and rentals, and computer processing. Not included in this figure are research and development costs or indirect costs to a participating institution.

A Test-Retest Study of the COMP/ACT Measurement Battery, Form II

A test-retest study of the COMP/ACT Measurement Battery was conducted in March 1978 at two institutions with those students who had participated in the November 1977 field test. Designed to permit evaluation of as many sources of measurement error as possible, the study aimed to determine the reliability of measurement materials in conditions approximating those of actual use.

One issue explored was the stability of student responses to open-ended items. The interval of 14-16 weeks between test dates was sufficient to expect learning to have occurred which could affect the relative ranking of students. Another issue, also related to open-ended items, concerned the reliability with which rating scales were applied to judge the adequacy of student responses. For the first test, faculty evaluators at the ACT National Office rated the responses, while for the retest, faculty evaluators at each of the two institutions did the rating. Ordinarily, reliability of raters is determined by comparing the degree of agreement of different judges rating the *same* set of student responses. Here ratings by one set of judges of the first test responses are compared to ratings by a second set of judges *of a second set of student responses*—the same students' narrative responses to a second administration of the Battery.

There were several other sources of potential measurement error, including possible variations in test administration conditions and instructions. Student motivation and effort to take this six-hour test twice may have affected performance. Student volunteers took the first test. On the retest, they were paid $3.00 per hour and encouraged to try for higher scores. They were not, however, given results of their first performance until after the retest.

At one institution 29 of the original 41 students (71%) participated in the retest. At the second, 41 of the original 46 students (89%) participated. Approximately the same proportion, and about equal numbers, of freshmen and seniors at each institution took the retest, but about twice as many women as men were represented. On each test student responses to each question were rated on a five-point scale; these ratings were then summed to produce a total score for the Measurement Battery. For each institution, total scores for the November test were correlated with retest total scores to determine whether the students ranked in the same order each time. At one institution this reliability coefficient was .90; at the other institution it was .89. Such coefficients are unusually high for open-ended assessments of this nature, where results over .70 are generally considered good. In view of the many sources of measurement error outlined above, they are remarkable. These results suggest that the Measurement Battery can be used to make judgments about individual students and that confidence can be placed in the reliability of the total assessment procedure.

Correlation coefficients for the Battery's six major area subscores were also quite high, as Table 4 indicates; at one institution all correlations were .78 or above; at the other institution, one correlation was .66 while the remaining coefficients ranged from .73 to .88. These high correlation coefficients are adequate for using and interpreting results at individual and group levels.

TABLE 4

Test-Retest Correlations of Measurement Battery Scores for Two Institutions

	Total Score	Functioning in Social Institutions	Using Science	Using the Arts	Communicating	Solving Problems	Clarifying Values
Institution 1	.90	.79	.78	.82	.80	.82	.84
Institution 2	.89	.73	.66	.75	.79	.88	.84

Correlations for the other subscores (except for Writing and Speaking) were generally at the .60 level and are adequate for using these subscores to evaluate programs. Subsequently, special effort was made to refine the procedures for evaluating writing and speaking skills. As described in the last section of this report, the resultant reliability coefficients for judgments made in these two areas have been increased to the .80 level.

Another feature noted in the test-retest study was variability in mean level of performance. At one institution mean scores remained quite stable. At the other, mean scores increased by about 25%, with seniors showing greater increases than freshmen. Further studies must determine whether this increase in level of performance was due to learning, to differences in test administration and rating, or to an educational experience within the test itself.

A Proxy Measure for the Measurement Battery: the COMP/ACT Objective Test, Form I

A study of the characteristics of the COMP/ACT Objective Test was conducted in March 1978 at five institutions, using students who participated in the November 1977 field test of the COMP/ACT Measurement Battery. The Objective Test is the product of an effort to create proxy measures for all components of the Measurement Battery. Because this field test constitutes the test's first use, it offers only preliminary evidence of usefulness as a proxy measure.

The Objective Test contains 63 multiple-choice questions requiring application of general knowledge and skill to realistic adult situations. Each question has two correct options and two incorrect options. Selection of an incorrect option by a student results in a point being subtracted both from the subscore to which the question is related and from the total score. The range of possible total scores on the test is from -126 to +126 (i.e., 252 points), and all scores are adjusted to reflect positive scores only. Both items and corresponding subscores are designed to parallel those of the Measurement Battery. Participant responses are recorded on machine-scorable answer sheets. Test administration requires two and one-half hours. As with the Measurement Battery, there are administration and scoring options to make the test modular.

Five institutions participated in the first field test of the Objective Test, with a total of 180 (76%) of those students who took the Measurement Battery in November 1977 participating in the study. The sample was about evenly divided among freshmen and seniors, males and females. Mean ACT Composite scores for four of the five institutions were roughly equivalent (24, 23, 24, 25). The mean ACT Composite score for the fifth institution was higher than for the other student groups (29).

This first use of the Objective Test suggests that it measures general education knowledge and skills accurately enough to use in program evaluation.

- As a proxy measure for the Battery, the Objective Test will be more useful for making judgments about groups than for individual students. The correlation between total scores of the Battery and Test was .60. Correlations between subscores for the two measures, in each of the six major areas, ranged from .37 to .48, as illustrated in Table 5.

TABLE 5

Correlations of Measurement Battery and Objective Test Scores

Total Score	Functioning in Social Institutions	Using Science	Using the Arts	Communicating	Solving Problems	Clarifying Values
.60	.42	.48	.37	.44	.42	.47

With further revisions in the Objective Test, it is anticipated that correlations in all six major areas will be raised above .60.

- As an independent measure of general education, the Objective Test total score has an estimated reliability of .85, using Kristof's estimate of reliability for a three-part test. The mean score for seniors is significantly higher than for freshmen. There were significant differences in all the major area mean subscores except in Communicating. However, Kristof reliability estimates for subscores are in the .50 to .70 range. With further revisions of the Objective Test, it is anticipated that all subscore reliabilities will be raised to acceptable levels.

It is not likely that acceptable proxy measures of writing and speaking will be developed. Thus, institutions which use the Objective Test may wish to use the Measurement Battery Writing and Speaking subtests to assess these areas. Inclusions of these two subtests in the Communicating subscore is almost certain to result in a scale that discriminates well.

- As with the Measurement Battery, there is nearly zero correlation between the Objective Test total score and age. This suggests that the abilities being measured are unrelated to simple maturation. Correlations with major area of interest and sex were almost zero.

- For the seniors who responded to both the Measurement Battery and the Objective Test, the correlation of total score with GPA is significant, but small (.27). This compares with the correlation between GPA and Measurement Battery total score (.35). The Objective Test thus appears to be assessing abilities not measured by college grades.

- The Objective Test appears to be more closely related to measures of general ability than the Measurement Battery. ACT Composite scores for students who took both tests yielded correlations of .52 for the Measurement Battery and .70 for the Objective Test. Nevertheless, half of the differences in performance on the Objective Test cannot be attributed to general ability.

Vocational-Technical Student Use of the COMP/ACT Measurement Battery

The purpose of this section is to present an analysis of scores obtained by 258 vocational-technical students who took a part of the COMP/ACT Measurement Battery during 1977-78.

In order to make an appropriate comparison between the performance of these vocational-technical students on the Battery and that of students attending 4-year colleges participating in COMP, samples from each group were matched in ability, age, and sex. Ability scores (Florida 12th grade total scores) were available for only 81 of the vocational-technical students and 31 of these scores were below the ability range of the college sample. Therefore, a random sample of 50 college seniors was selected, and matched with the 50 vocational-technical students in ability, age, and sex. A table of concordance was used to equate the Florida 12th grade total ability scores with ACT composite scores for college seniors. The two samples thus matched were of the same size (N = 50 each). The average age of each sample was identical (23 years), and the age range was also roughly equivalent (19-37 years for the vocational-technical group and 20-34 years for the college seniors). At the time the two groups were seniors in high school, they had reached roughly equivalent levels of general academic development, based on an average ACT Composite score of 20.36 for the college group and a Florida 12th Grade Test score average for the vocational-technical students equivalent to an ACT Composite score of 18.94. The ACT Composite score range for the college sample was 12-26 and the range for the vocational-technical sample was 10-26.

The two samples had roughly the same number of males and females, with the college sample being 44% male and the vocational-technical sample 46% male. The college sample was drawn from six different colleges participating in COMP and the vocational-technical sample represented eight of the Florida vocational-technical centers. Given these efforts to match ability, sex, and age, the interpretations which follow are based on the assumption that the two samples are comparable.

All 15 sets of stimulus materials and questions in the Battery were administered to college students at colleges participating in COMP field tests, whereas only 12 of these sets were given to the vocational-technical students. To make appropriate comparisons, scores were recomputed using the same 12 sets of responses for the 50 college seniors in the matched comparison group. The average scores obtained by the two samples are displayed in Table 6.

TABLE 6

Mean Scores for the COMP/ACT Measurement Battery (Modified Form)

	Total Score	Functioning in Social Institutions	Using Science	Using the Arts	Communicating	Solving Problems	Clarifying Values
College Seniors (N = 50)	44.52	7.70	5.25	6.94	8.10	10.14	6.36
Matched Group of Voc/Tech Students (N = 50)	27.22	4.52	2.00	3.62	6.56	7.02	3.50
Total Group of Voc/Tech Students (N = 258)	23.81	3.86	1.76	3.19	6.00	6.30	2.70

The college seniors obtained mean scores substantially above those of the vocational-technical students. Furthermore, the vocational-technical students responded to a modified version of the Measurement Battery, omitting activities 5, 10, and 15, which constitute over one-third of the questions judged to be the most difficult and complex in the Battery. Had the vocational-technical students responded to the total Measurement Battery, the differences in mean scores might have been even greater. Since the two groups were judged to have the same level of general educational development at the end of high school, we tentatively conclude that the college experience is responsible for further development of the college seniors. Yet because a number of factors influence educational development, these results merely suggest developmental differences resulting from different educational programs. Further research on this issue is planned for the future.

One further interesting bit of information appears in the data of this study. Assuming that a total score at the 25th percentile of the college group is commensurate with the general education competency expected for the awarding of a B.A. degree, 21% of the 258 vocational-technical students could be judged as already meeting the intent of many undergraduate general education programs. Table 7 develops this data further.

TABLE 7

Percentage of Vocational/Technical Students Scoring at or above the 25th Percentile for the Comparison Group of College Seniors

	Total Score	Functioning in Social Institutions	Using Science	Using the Arts	Communicating	Solving Problems	Clarifying Values
Raw Score at 25th %tile	32	6	3	4	6	7	4
Matched Group of Voc/Tech Students (N = 50)	32%	34%	28%	44%	72%	60%	30%
Total Group of Voc/Tech Students (N = 258)	21%	24%	26%	38%	59%	45%	26%

Since the vocational-technical students did not take the entire Battery and these 258 students may not be representative of all vocational-technical students, it is difficult to estimate how many such students may already possess general education skills and knowledge, as measured by the Battery, at a collegiate level. However, the evidence provided by this study suggests that many such students may exist. Should these students wish to transfer to a program leading to a liberal arts degree, it would appear appropriate for them to receive credit for these skills.

Mature Adult Performance on COMP/ACT Instruments

Since a major goal in developing the three COMP instruments has been to measure the ability of persons to function effectively in adult society, it seems appropriate to examine the performance of adults on the Measurement Battery, Objective Test, and Activity Inventory. Such investigation of adult performance should provide an indication of the predictive validity of the instruments for effective functioning in adult roles. Of three such studies conducted in 1977-78, two involved 58 adults in Milwaukee and 30 adults in Miami. The third involved 41 college students at least 25 years old enrolled in five of the institutions participating in COMP. These studies were of an exploratory nature and were designed to gain a very general idea of the effectiveness of the three COMP instruments.

The adults selected to participate were recommended by leaders of community volunteer organizations as those effectively functioning at an amateur level in one of the following roles: clarifying social values, solving technological problems, or communicating about the arts. The adults were asked to select one of the following areas as being of greatest interest to them: social science, natural science, or arts/humanities. In 80% of the cases, the area of greatest interest selected by the participant matched the role for which that individual was identified. Thus, the roles which most interested the participants were also, by implication, those for which they were best qualified. The older college students were also asked to identify which area most interested them so that the following comparison could be made.

All participants completed the open-response version of the Activity Inventory, which was then scored using rating scales developed in the project. These adults definitely tended to score highest on the section of the Inventory that corresponded to their self-reported area of greatest interest (statistically significant above the .001 level with a contingency coefficient of .67 out of a possible .82). Furthermore, the mean score for those identified with social science on the Functioning within Social Institutions (FSI) section of the Inventory was higher than the mean scores for those identified with natural science and arts/humanities. Those identified with natural science had a higher mean score in Using Science and Technology than did those in social sciences and arts/humanities. Those identified with arts/humanities had a higher mean score in Using the Arts than did those in the social and natural sciences. Findings from these initial studies offer evidence of the validity of the Activity Inventory in identifying effectively functioning adults.

As a result of this evidence, three or four greatly expanded studies of the validity of the Activity Inventory are planned for 1978-79. Also, the responses provided by the two adult studies, together with those provided by 470 students, have aided in developing an objective version of the Inventory as well as the rating scales used to score the open-response version.

The next step was to determine whether scores on the Measurement Battery obtained by adolescent students might be predictive of future functioning in adult roles, by examining the correlation between Battery and Activity Inventory scores. The Milwaukee and Miami adults responded to only one-third of the Battery. Even so, the total scores for this partial Battery showed rather high correlation ($r = .43$) with Inventory total scores for these adults. Again, this evidence should encourage further studies of adult performance on the Battery during 1978-79. In the study of 41 students over the age of 25 at five COMP participating colleges, a correlation of .69 was obtained when these adults took the entire Battery as well as the entire Inventory.

The third step in this study was to determine the degree of correlation between scores on the Objective Test and the Battery. Not only would this yield information about the degree to which the Test could be used to predict scores on the Battery, but it would indicate how well the Objective Test might serve as an inexpensive proxy measure in evaluating the general or liberal education program of a college. Results for the three groups varied widely. For the Milwaukee adults, a correlation of .46 was obtained between Test total scores and Battery total scores. For the Miami adults, the correlation was .59. Here again, it is important to remember that these adults took only one-third of the Battery. In the case of the 41 older students, the 17 who took both the entire Battery and Objective Test showed a correlation between total test scores of .74. Although variable, these correlations suggest it is possible to generate viable proxy measures. ACT staff are greatly encouraged and will continue development of the Objective Test in 1978-79.

Improved Reliability of Rating Speaking and Writing Skills

Speaking and Writing are abilities held to be extremely important for effective functioning in a variety of adult roles. Yet existing measures of these skills at the college level do not generally result in high levels of agreement between raters. Furthermore, the existing measures are often more like activities students do in classrooms than tasks adults must do in their daily lives. Hence, their validity as accurately reflecting abilities important to effective functioning is also open to question.

As part of the College Outcome Measures Project, direct measures of Speaking and Writing have been developed that use realistic tasks to measure applied college exit level skills. These are included as part of the COMP/ACT Measurement Battery.

In two test-retest studies of this Battery, total test score reliabilities of .90 and .89, and subscore reliabilities of .66 to .88, were obtained. Reliability coefficients for the Communicating subscore were .80 and .79. When two components of this subscore, Speaking and Writing, were examined, however, reliability coefficients were found to range from .37 to .72. The studies summarized in this section were initiated to determine whether improvements in rating procedures plus training of evaluators would raise interrater agreement to acceptable levels.

Using data from the test-retest study of the COMP/ACT Measurement Battery mentioned above, two trained raters independently rated 74 sets of oral responses and 65 sets of written responses. In the Speaking subtest, composed of role-playing tasks in each of the three content areas of social sciences, natural sciences, and the arts, students were given 24 hours before audio- or videotaping three 3-minute responses. The Writing subtest involved three role-playing tasks using taped stimuli in the same content areas, requiring 60 minutes of writing to compose three letters directed to various audiences.

Raters were required to apply qualitative rating scales, defining five levels of proficiency on three broad criteria, to each of the three oral or written responses. Guidelines were provided to minimize various sources of bias in rating. As a training procedure, the two raters involved in the Speaking and Writing studies were given practice in rating ten student responses to each of the three writing or speaking tasks. They were then asked to discuss their ratings with each other and justify any discrepant ratings. Finally, they were given feedback on how others had rated the responses. Differences in level as well as spread of rating were pointed out and strategies to compensate were discussed. This practice rating session required about six hours for each of the two studies.

Two approaches to studying interrater agreement were used. First, as the five-point rating represented continuous data, a Pearson product-moment correlation was calculated between rater 1 and 2 for the combined ratings of all three student responses. Second, the percentage of total agreement and agreement within one category were calculated for ratings of all three student responses combined. Percentages of agreement on a pass/fail basis at each level of proficiency were also calculated.

For the Speaking subtest, the correlation between two independent ratings by trained raters was .83. Percentage of **exact** agreement about which of the five levels of proficiency each student response should receive was 51% for trained raters, or double the figure for untrained raters. Trained raters achieved 94% agreement on ratings within one category on the rating scale. This resulted in extremely accurate ratings of student responses at minimally acceptable and advanced levels of proficiency. When evaluating speakers on a pass/fail basis, trained raters agreed 99% of the time about whether performance fell above or below the lowest standard. At the second level of proficiency, 87% agreement was achieved; at the third level, 77%; and at the highest standard, 83%. Thus, it appears that Speaking can be objectively and reliably rated by trained raters using COMP/ACT materials and rating procedures. The validity of such measures in relation to effective functioning in adult roles is currently being studied.

For the Writing subtest, the correlation between two independent ratings by trained raters was .75. Percentage of **exact** agreement about which of the five levels of proficiency each student response should receive was 35%. Trained raters achieved 84% agreement within one category on the rating scale, and the accuracy of rating student responses at minimally acceptable and advanced levels of proficiency approached that of the Speaking study. When evaluating Writing on a pass/fail basis, trained raters agreed 91% of the time about whether performance fell above or below the lowest standard. At the second level of proficiency, 72% agreement was achieved; at the third level, 71%; and at the highest standard, 82%. This pattern suggests that the lower correlation obtained in the Writing study is due to disagreements at intermediate levels of proficiency. With further refinements in the rating scale and training materials, it is likely that even higher levels of interrater agreement can be obtained.

These studies demonstrate that Speaking and Writing can be measured reliably. Using such measures, normative data can be generated for college freshmen and seniors. These data will be highly relevant to the current emphasis on basic skills at the public school level.

A more immediate and practical use for these measures is in examining and improving college programs which aim to develop Speaking and Writing skills. Many postsecondary institutions are currently reexamining their general education programs and the meaning of a liberal arts degree, and these data might aid in studying alternative instructional strategies. Using such measures, students could also be screened and more appropriately placed in programs addressing their needs. For institutions wishing to verify the development of Communication skills at acceptable levels, measures such as these could be used to assess growth and certify competence. Finally, the assessment and rating procedures offer a model for college faculty to use in developing and reliably rating other open-ended measures of skills and knowledge.

REPORT OF THE COLLEGE OUTCOME MEASURES PROJECT ADVISORY AND EVALUATION PANEL FOR THE PERIOD SEPTEMBER 1976 THROUGH AUGUST 1978

Morris T. Keeton

The COMP Advisory and Evaluation Panel is composed of Dr. George L. Hall, Executive Director, State Board of Directors of Community Colleges of Arizona; Dr. Wayne H. Holtzman, President, the Hogg Foundation for Mental Health; Dr. Morris T. Keeton, Executive Director, Council for the Advancement of Experiential Learning; and Dr. Patricia A. Thrash, Associate Director, North Central Association of Colleges and Schools Commission on Institutions of Higher Education.

The Advisory Panel of the COMP/ACT project, funded by the Fund for the Improvement of Postsecondary Education, has asked me to report to you on the panel's evaluation of the project to date. In doing so, I have the benefit of written comments from the other members of the panel as well as the experience of our meetings and the documents produced during the project.

The purpose of this project was to generate a model solution to a most important problem in higher education: the lack of a valid, reliable, and cost effective process (with implementing tools) for assessing results on a well-conceived and defined set of general education outcomes of college education (baccalaureate level). The desired outcomes were seen, not simply as knowledge acquisition nor simply as general academic aptitude, but as the capabilities which effectively functioning adults would need for a personally rewarding and socially useful life in this society. It was a high risk endeavor to seek to generate such a process, and the risk was compounded by the difficulties of assuring acceptance and use of the model by institutions of higher education.

The purpose, high risk though it was, has nevertheless been achieved in the judgment of the panel. The dissemination of uses of the model beyond the experimental group of institutions still lies ahead; but the most difficult hurdles—developing a suitable definition of general education outcomes, enlisting an appropriate array of experimental institutions, achieving reliability in the measures, achieving valid measures, and doing these things in a way that would promise cost effectiveness—have been overcome. The basis for this judgment is here presented in the form of answers to four questions:

1) Do the outcomes selected for assessment meet the tests appropriate to the purpose of the project? a) Are they learnable through baccalaureate programs of general or liberal education? and b) Do they give promise of appropriate correlation with effective functioning in adult society?

2) Is it likely that the three assessment instruments developed in the project will prove to be valid, reliable, and cost effective?

3) How effective have the dissemination efforts been? a) Are the communications clear, relevant to both the purpose of the project and the likely concerns of user institutions, and thorough in dealing with the available information? b) Have the dissemination efforts been appropriate to the purpose of eliciting eventual wider use of the project outcomes?

4) Do the probable benefits of the project render the outlay of funds cost effective?

1. *Appropriateness of the Selection of Outcomes to be Measured.*

A number of elements enter into the definition of what is appropriate as a selection of outcomes to be measured in this project: a) For traditional definitions of what the outcomes of general education should be, there is no need for new measures. Reliable ones now exist. While the selection of particular test items might be questioned, there are already arrangements in being for periodic review. The relevance of these traditional definitions to what is needed by the society from college graduates is also under increasing question. The selection of outcomes to be measured in this project therefore needed to reflect a different concept of the intent of general education. b) The most promising line of inquiry or exploration appears to be one which speaks to the capabilities achieved by students rather than the particular knowledge content mastered. But what capabilities? Traditional measures have correlated best with success in later college studies, much less well with effectiveness in adult roles in work. Little has been learned about the predictability of effective functioning of college graduates in other later life roles. The decision to try to define the capabilities of "an effectively functioning adult" and to develop outcome measures with significant predictive value for such functioning was therefore one likely to be useful to both the institutions of higher education and their graduates. c) The ideal choice of outcomes would be one which made possible measurements which could both facilitate program evaluation and improve the usefulness of assessment information to both the students (graduates) and their employers and peers. d) The outcomes to be measured should differ significantly from related and already measurable changes that occur during the course of college years; e.g., from simple maturation, from academic aptitudes as defined in existing measures, and from achievement or knowledge tests such as those in the Graduate Record Examination series.

The six outcome "areas" chosen for measurement consist of three capability areas (communications capabilities, problem solving capabilities, and value clarification capabilities) interacting with three arenas for application (functioning within social institutions, using science and technology, and using the arts) to yield nine areas of ability to function in adult roles. These outcome areas seem to the panel an appropriate choice using the criteria just stated, and the findings of the field tests of the COMP/ACT Measurement Battery support the judgment that these outcomes can be, and are, affected by college education which is designed to have those effects and do correlate significantly with the characteristics of effectively functioning adults.

- Spring and fall field tests were conducted in 1977 on Forms 1 and 2 respectively of the Battery. "The results of the first field test show a clear progression of growth across four years of college. . . . The results of the fall field test for a new form of the Battery show highly significant differences between freshman and senior means. . . . In both field tests there is a nearly zero correlation between age and total test score. No significant differences in total test score were noted by sex or major field."[1]

[1]Material quoted is excerpted from an earlier version of the first study described in this report, "Fall 1977 Field Trials of the COMP/ACT Measurement Battery, Form II," originally circulated to participants in the project.

- A later 1977-78 study compared the performance of graduates of post-secondary vocational-technical programs with that of baccalaureate graduates of the project colleges on the COMP/ACT Battery. (See the fourth study described in this report, "Vocational-Technical Student Use of the COMP/ACT Measurement Battery," for details.) The baccalaureate graduates scored substantially higher than did the vocational-technical ones. Panel members, however, insofar as their comments have been received on this study, do not regard it as well designed to provide the kind of matching groups of students required for a decisive evaluation nor was the control of other factors strong. There is need for the time and other conditions that will permit pre- and post-training measures and other steps to reduce the possibilities that artifacts will enter into the results.

It will always be possible to debate as to what outcomes colleges should be trying to generate in their general education programs, what priorities the different objectives should have, which of those such as here addressed can be "taught deliberately," and like questions. In this instance a sizeable group of colleges, which are known to represent interests shared by a larger company, had the opportunity to try to clarify what they most valued among the possible outcomes to be sought and assessed. Given the current challenge as to whether college education is really worthwhile and the extensive pressures to answer that question primarily in terms of job- and career-related concerns, the choice made by these colleges and applied in this project is a significant effort to clarify and revitalize the commitment to complementary values. It reaffirms the idea that college education should be judged by the quality of life to which it contributes and that this quality is a matter of the development of interests and concerns and of behavior applying those concerns in participation in the adult community.

2. *The Validity, Reliability, and Cost Effectiveness of the Assessment Instruments and Processes.*

The most appropriate general answer to the questions on the technical adequacy of the COMP/ACT Battery is that it has achieved reasonable validity, reliability, and cost effectiveness.

The validity of such instruments can be at best only relative. To obtain a strong confirmation that the results of use of this Battery will measure what really makes for effective functioning by adults is a formidable undertaking. It cannot possibly be completed until enough time has passed to have the early examinees grow into such adults and undergo further testing. The project has, since time must pass before such testing can be done, chosen intermediate clues and criteria which the panel views as reasonable ones. The face validity of the measures is good. We have not been able to suggest better ones. Certain findings to date are reassuring. For example, the Battery does not measure some things it was meant not to measure. It is not simply measuring college grades. The correlation between age and total test scores is near zero (the Battery is not simply measuring maturation). The Battery

measures something other than general aptitude and ability (the correlation leaves over half of the variance in performance unexplained on the total scores). On the positive side, the studies done with older adults identified as effective in the areas under study showed a statistically significant correlation between high scoring on the corresponding tests and the identified areas of their greatest effectiveness. Cultural bias does not appear to be working in the results.

With regard to the reliability of the tests, three kinds of reliability appear to have been achieved at suitable levels. Interjudge agreement in the ratings is very high. On internal consistency and homogeneity of scores, the evidence is not as extensive, but what there is looks good. On test-retest studies the reliabilities are quite high. For such an early stage of test development, these are quite heartening results. (See the second study described in this report, "A Test-Retest Study of the COMP/ACT Measurement Battery, Form II," for details.)

In evaluating the reliability of the instruments and the assessment processes careful attention must be given to the purposes for which the results are to be used. It must also be remembered that the Battery is still in the process of being refined and that further development of both the instrument and the training of judges may result in improved reliabilities. In the fall 1977 study of the Battery seven purposes are identified for which the results might be used. Of that of program evaluation and planning the Battery already provides quite adequate levels of reliability. The remaining functions have to do with predicting individual performance and admitting people to opportunities or placing them within programs or enterprises. On these functions a distinction should be made between functions having to do with selection among competing candidates and those having to do with guidance or counseling for the individual. The reliabilities of the scores obtained with the Battery are already approaching levels adequate for the selection functions, particularly in view of the fact that it is difficult to find alternative means that approach such levels of reliability. On the functions related to guidance, there is a considerable way yet to go before the Battery scores can be regarded as adequate.

The studies and reports made on the Battery to date give a clear picture of progression from the first rough-hewn construction efforts to the present, more refined condition of the instruments. The trend gives a strong sense that the task of reaching adequate sophistication in the instruments can be done. The researchers have made impressive efforts to expose the possible sources of error in rating. They have been candid in reporting needed improvements as well as encouraging findings. The likelihood of locating problems and removing them therefore seems strong.

The panel has been concerned that measures developed in this project should begin to get at the higher levels of intellectual performance and problem-solving skills as defined, for example, in the Bloom and Broder taxonomy of cognitive objectives. This task is a difficult and vexing one. We feel that the project staff has made some headway on it. The Activity Inventory is one expression of this effort, may prove to be only a beginning, but may also prove useful as a "stand-alone" instrument. This whole effort

may make a contribution to the measurement of the contributions of experiential learning to college education. We also see the effort, after achieving a reliable measure of a different kind of college outcome from those traditionally measured by objective tests, to find an Objective Test which will serve as a proxy measure for this new type of outcome as a worthwhile one.

As to the cost effectiveness of the instruments and processes, the panel had had grave misgivings early in the project. We are impressed with the progress made on this concern. The cost estimate of some $25 per student in the most recent study appears to be accurately developed. When one considers the $2500-$7000 per year spent in the general education of students, and notes that up to the full-time equivalent of two years may be devoted to the general education requirements of a college, an expenditure of $25 per student to assess the outcomes appears very small indeed. Further refinement of the instruments and procedures is warranted; and if it should prove twice or more times as costly to obtain results reliable for individual guidance, the cost would be well justified. The costs of such assessment in portfolio development and evaluation, though yielding more individualized and detailed specification of learning outcomes, is so much greater (comparable to the cost of an entire course or two) that the COMP/ACT Battery, alone or in combination with such other means, would be a significant contribution to quality assurance and to effective educational planning and guidance.

In connection with both the questions of reliability and of cost effectiveness, it should be noted that failure of some of the participating institutions to enlist sufficient numbers of subjects has yielded numbers insufficient to permit useful interinstitutional comparisons on achievement levels on the learning outcomes measured by the COMP/ACT Battery. Ultimately it would be highly useful to develop information suitable for such studies.

3. *Effectiveness of Dissemination Efforts.*

Insofar as the question of effectiveness of dissemination efforts refers to the objective of spreading the use of the COMP/ACT Battery and affecting the design of curricula to reflect objectives of the kinds measured by the Battery, it is premature to try to evaluate this effectiveness. The effort to enlist this wider participation would not have been appropriate until the recently completed studies of validity and reliability were completed. That work is accordingly just now beginning.

It is nevertheless possible to comment on communications and dissemination efforts which have laid the groundwork for effective dissemination of the kind envisaged. The panel has seen communications addressed to participating institutions, to the panel itself, and to others seeking to understand what is going on even before the opportunity for participation arises. In general, these communications have shown clarity, thoroughness, relevance to their functions, and good applicability to the needs of the addressees. Both participant groups and the Advisory Panel have been, at the outset, provided clear

statements of their respective functions and of the purposes of particular meetings and stages of work. In addition, there has been thorough reporting after meetings as to both what happened in the meetings and what follow up activities had occurred or were expected. The meetings themselves have been efficiently conducted.

Dr. Patricia Thrash states well the experience of the Advisory Panel with the staff work in these terms: "The meetings were focused; material was relevant and applicable. The 'hands on' experience of the December 2-3, 1976 meeting, as well as the opportunity to meet participants, was an excellent introduction. The March 1-2, 1977 meeting between the Advisory Panel and the ACT staff provided an excellent opportunity to review the progress of the project and to hear projections for the future. The September 9, 1977 meeting of the Advisory Panel and project staff brought panel members up to date on plans for field testing a new measurement battery. The written materials provided since that time (and especially in May and August) have been clear and thorough. Even without face-to-face meeting, I feel fully apprised of the project's progress."[2]

The project has been widely publicized, explained, and discussed among constituencies likely to contribute to the success of later dissemination efforts. Such occasions as the annual meeting of the North Central Association (1978), the assemblies of the Council for the Advancement of Experiential Learning (twice in 1977), and a meeting of the Council on Postsecondary Accreditation (1978) have been used to spread the word of the project. The project was treated by Aubrey Forrest in "Competency-Based Assessment in Postsecondary Education—Some Issues and Answers," in the NCA Quarterly (Fall, 1977, Vol. 52, No. 2, pp. 322-326). The project, in summary, is widely known. Its use of participating institutions from the beginning throughout development has contributed both to the clarity of communications and to the acceptance and the likely receptivity to use of the battery as opportunity is opened to additional institutions.

4. *Cost Effectiveness of the Project.*

Paraphrasing Dr. George L. Hall of the Panel, COMP was a risky project. There was no certainty nor any guarantee that valid and reliable evaluation tools would result. It seemed almost too much to expect even that the several, vastly different participating institutions could collectively accept beginning definitions and measures of outcomes. Hence, members of the panel had some skepticism about the project in early meetings with the staff and institutional representatives. This skepticism has gradually changed to belief. It is the panel's thinking now that the major difficulties of the project have been overcome, that field tests are very satisfactory, and that the next year is most important in the dissemination and further development of the project.

[2]Letter from Patricia Thrash to Morris Keeton, September 4, 1978.

Dr. Thrash expresses this same judgment in this language: "All parties involved in postsecondary education need accurate information about general and liberal education outcomes in order to make sound educational decisions. This skill in assessing outcomes is still in its infancy; there is much more talk than evidence that it can be done. In this project, for the first time, a serious and workable attempt has been made to do what all of us—educators, students, accreditors, financial supporters—say needs to be done. The instruments thus far designed are cost-effective both in time and money terms. They are complementary. The promised continued commitment and support of ACT should serve as a guarantee to FIPSE that these instruments will find their way into the mainstream of postsecondary education. They will be of great assistance to individuals, at all stages of their development from college onward. They will be of extraordinary value to postsecondary institutions. I was impressed by the specific primary uses cited in the Fall 1977 Field Test Report. Each of the seven speaks to a concrete and urgent area of need (general or liberal education program evaluation and planning; academic advising; academic placement; credit to adults for experiential learning; certifying students for graduation; screening students for employment; and screening students for graduate or professional training). . . . The project involves a **timely** effort in a significant. . . area of education. I believe. . . that 'literally hundreds of postsecondary institutions would enthusiastically take advantage of the assessment instruments and procedures that result from this project.' I know of at least 200 North Central institutions that would be receptive to something as concrete and enabling as these instruments are.

"At its best, the COMP/ACT project represents a significant breakthrough at the best possible moment. If subsequent development and broader exposure of the instruments yield less sanguine results, a significant step forward will still have been made. I am not a researcher; I am an educational practitioner. My own experience tells me that students and institutions need these instruments badly. The evidence of their effectiveness is compelling."[3]

The panel wishes also to record its appreciation for a dedicated, hard-working, persistent, and competent staff and core of institutional representatives. The Fund for the Improvement of Postsecondary Education is fortunate to have entrusted this difficult venture to such good hands.

[3]Letter from Patricia Thrash to Morris Keeton, September 4, 1978.

ACT has been assisted in the second year of COMP by:

Alverno College
Brigham Young University
Colgate University
College of Du Page
Delaware County Community College
Florida A&M University
Florida International University
Lindsey-Hopkins Technical Education Center
Lively Vocational-Technical Center
Mars Hill College
Memphis State University
Miami-Dade Community College
Michigan State University
Mid-Florida Technical Institute
Okaloosa-Walton Junior College
Our Lady of the Lake University
St. Louis University
St. Petersburg Junior College
Sarasota Vocational-Technical Center
Seminole Community College
South Brunswick High School
State University System of Florida
Tennessee Higher Education Commission
Tennessee Technological University
University of Nebraska-Lincoln
University of Windsor
Washington-Holmes Vocational-Technical Center